TM94-2012
0262
Copy 1

NUWC-NPT TM 942012

NAVAL UNDERSEA WARFARE DIVISION
NEWPORT, RHODE ISLAND

Technical Memorandum

ROBUST TRAINING OF THE QUADRATIC CLASSIFIER

Date: 2 February 1994

Prepared by: Paul R Kersten
Paul R. Kersten
Technology & Advanced
Systems Division
Combat Control
Systems Department

Approved for public release, distribution is unlimited.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **02 FEB 1994**	2. REPORT TYPE **Technical Memo**	3. DATES COVERED **02-02-1994 to 02-02-1994**

4. TITLE AND SUBTITLE **Robust Training of the Quadratic Classifier**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **Paul Kersten**	5d. PROJECT NUMBER **802424**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Undersea Warfare Center Division,1176 Howell Street,Newport,RI,02841**	8. PERFORMING ORGANIZATION REPORT NUMBER **TM 942012**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **Office of Naval Research**	10. SPONSOR/MONITOR'S ACRONYM(S) **ONR**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
NUWC2015

14. ABSTRACT

The quadratic classifier is one of the most applied parametric classifiers used in pattern recognition. To use this classifier, one trains it by estimating the center and the dispersion of the different classes from the data. These estimates are usually made using sample means and sample covariances. If the data errors are normal, this is the optimal procedure. However, in practical situations where the data are not normal or contain outliers, the training can fail because the estimation procedure is not robust. This technical memorandum describes a robust method of estimating these parameters. This estimation method is much more resistant to outliers and perturbations from the assumed normal distribution than existing methods.

15. SUBJECT TERMS
Fuzzy Expert Systems; quadratic classifier

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**36**	

ABSTRACT

The quadratic classifier is one of the most applied parametric classifiers used in pattern recognition. To use this classifier, one trains it by estimating the center and the dispersion of the different classes from the data. These estimates are usually made using sample means and sample covariances. If the data errors are normal, this is the optimal procedure. However, in practical situations where the data are not normal or contain outliers, the training can fail because the estimation procedure is not robust. This technical memorandum describes a robust method of estimating these parameters. This estimation method is much more resistant to outliers and perturbations from the assumed normal distribution than existing methods.

ADMINISTRATIVE INFORMATION

This report was funded under NUWC Newport IR/IED Project No. 802424 "Fuzzy Expert Systems." The IR/IED program is funded by the Office of Naval Research; the NUWC Newport program manager is K. M. Lima (Code 102).

The technical reviewer for this memo was S. E. Maloney (Code 2211).

ACKNOWLEDGMENTS

The author gratefully acknowledges the enthusiastic support of D. L. Reade and O. McNiel of the Target Recognition Branch of the Naval Air Warfare Center, Weapons Division, China Lake, CA.

TABLE OF CONTENTS

Section **Page**

LIST OF ILLUSTRATIONS

LIST OF TABLES

ROBUST TRAINING OF THE QUADRATIC CLASSIFIER

1. INTRODUCTION

This technical memorandum describes the robust training of the quadratic classifier (QC). The QC is derived by assuming a normal distribution on the data and applying the maximum likelihood principle (reference 1). Traditionally, the sample mean and covariance are used to estimate the two sets of parameters needed to construct the classifier. When the training data are contaminated or do not have a normal distribution, these estimates are no longer optimum and in fact yield erroneous results. It is a known fact that the sample mean and the sample covariance are vulnerable to outliers. The approach developed here uses robust estimates of the centering vector and the covariance matrix that are resistant to outliers and heavy-tailed data.

To evaluate the QC for the two-class problem, the centering vector and the covariance matrix must be specified for each class. Estimating these parameters from the training data is called training the QC. Two examples illustrate why robust training of the QC is necessary. Both illustrations are based on changes in the underlying modeling assumptions in the data. The first replaces the normal data assumption with contaminated normal data, and the second replaces the normal data assumption with the Cauchy distribution model. In both cases, the traditional classifier performance deteriorates but the robust classifier remains stable. Therefore, when training a QC from real data it is *not* a good idea to use the sample means and covariances. A better idea is to apply robust statistics that work just as well as sample statistics when the normal assumption is satisfied and, in addition, degrade gracefully when the training data are contaminated or are non-normal in nature.

The QC is defined and illustrated in section 2 on normal clusters. Since the classifier is designed assuming normality, this section illustrates the best performance of the classifier. In section 3, two examples are given of non-normal data and it is shown how the classifier design deteriorates. Two distinct sources of outliers are illustrated: mistakes represented as a clump of outliers contaminating the data and mismodeling represented as heavy-tailed distributions instead of normal random variates. In section 4, robust statistics are discussed that can make the training of the classifier more resistant to outliers. In section 5, the same two examples are illustrated when the training data are used to estimate the parameters of the QC using robust statistics. Section 6 consists of the summary and the conclusions.

2. QUADRATIC CLASSIFIER

2.1 BACKGROUND

The QC is the maximum likelihood classifier for the two-class problem with an underlying multivariate normal distribution where, in general, the covariances and the means of the two clusters are different (reference 1, p. 16). The test statistic associated with the QC (reference 1, p. 54) is defined for the multivariate normal as $h(\boldsymbol{x}) = -\log(l(\boldsymbol{x})) = \log$ *likelihood* where

$$h(x) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_1)^t \Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{m}_1) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_2)^t \Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{m}_2) + \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|}. \quad (1)$$

Here $\boldsymbol{m}_1, \boldsymbol{m}_2,\ \Sigma_1, \Sigma_2$ are the true first and second order central moments of the two classes. $P(\omega_1)$ and $P(\omega_2)$ are the *a priori* probabilities of a sample being from class 1 or from class 2. Then the decision rule is given by: Choose class 2 if $h(\boldsymbol{x}) > \frac{1}{2}\log\frac{P(\omega_1)}{P(\omega_2)}$; otherwise, choose class 1. Here, $\boldsymbol{x}^t = [X_1, \ldots, X_d]$ is a random vector or a vector where each component is itself a random variable. If $\Sigma_1 = \Sigma_2 = I$, then this equation reduces to the linear classifier given by $h(\boldsymbol{x}) = \boldsymbol{x}^t(\boldsymbol{m}_2 - \boldsymbol{m}_1) + \frac{1}{2}(\boldsymbol{m}_1^t\boldsymbol{m}_1 - \boldsymbol{m}_2^t\boldsymbol{m}_2)$. Usually, sample statistics are used to estimate $\boldsymbol{m}_1, \boldsymbol{m}_2$ and Σ_1, Σ_2. Often, the decision rule is represented using discriminant functions $g_i(\boldsymbol{x})$, where

$$g_i(x) = -\frac{1}{2}(x - \boldsymbol{m}_i)^t \Sigma_i^{-1}(x - \boldsymbol{m}_i) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\log P(\omega_i) - \frac{d}{2}\log 2\pi,$$

$i = 1, \ldots, c$ and d is the dimension of the vectors (reference 2, p. 17). Then one decides class i if, $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})$ for all $j \neq i$. These approaches for the decision rule are equivalent.

Here robust means gross error sensitivity, which intuitively is the worst influence one point can have on the statistic (reference 3, p. 87). The sample mean and covariance are defined by the following two equations:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \text{ and } S_N = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^t.$$

Neither of these statistics is robust; that is, one outlier can destroy the statistic. For the sample mean $\bar{x}$, this is apparent because if any element in the sum, say x_i, is extremely large in magnitude, it will dominate the finite sum and thus the statistic will break down, i.e., no longer be a reliable measure of central tendency. If $\bar{x}$ breaks down, then so does S_N, since all the components will be centered incorrectly causing the dispersion to totally break down. One measure of a statistic to withstand outliers and remain a reliable estimator is its breakdown point ε^*. Roughly speaking, this is the largest percentage of outliers that a statistic can handle in a sample without breaking down (reference 3, p. 97). The sample mean has $\varepsilon^* = 0$ (reference 3, p. 99) and thus so does the sample variance. So sample statistics are not robust; however, these statistics are optimal since they are the uniformly most powerful unbiased estimators of these parameters if the sample is truly normal.

The QC is well known and has several advantages. First, the QC is easy to visualize. The decision regions in two dimensions are conics and in higher dimensions are hyperquadrics. Intuitively, this allows one to think of one class being separated by a set of planes, sheets of a hyperboloid, or an ellipsoid. Second, the quadratic form separates the effects of location and scale into distinct terms. The first two terms of $h(x)$ (equation (1)) result from the difference in means. The third term of $h(x)$ results from the relative shape and scale of the two class distributions. The third advantage is that the QC is well known and easy to implement. Moreover, once reasonable estimates of the covariance and mean are obtained, the data can be squared-up or renormalized by simultaneous diagonalization (reference 1, p.31). Finally, the QC is optimal for the multivariate normal, which may be a good assumption for the center of a class of distributions from which the data are sampled. So in this memorandum the form of the classifier is retained. The training of the classifier is "robustified" and made resistant to changes in the underlying distribution and to data aberrations such as outliers. This resistance will be shown in later sections. In the rest of this section, examples of the QC are given.

2.2 QUADRATIC CLASSIFIER APPLIED TO NORMAL DATA

Since the quadratic classifier applies to multivariate normal distributions, one classic example is the case of antipodal signals with $\Sigma_1 = \Sigma_2 = I$ and $\boldsymbol{m}_1 = -\boldsymbol{m}_2$ with $P(\omega_1) = P(\omega_2) = \frac{1}{2}$ so the decision boundary is the line $x = 0$. Points on the left-hand-side (LHS) of the y-axis are in class 2 and points on the right-hand-side (RHS) are in class 1. This is illustrated in figure 1, with $\boldsymbol{m}_1 = [1.27,0,0,0,0]^t$. The sample consists of 400 points from a normal multivariate distribution, 200 points from a $N(\boldsymbol{m}_1,I)$, and 200 points from a $N(\boldsymbol{m}_2,I)$ and only the first two axes are displayed. The "o's" are from class 2 and the "x's" are from class 1. The discriminant function evaluated at the threshold is $h(\boldsymbol{x}) = 0$ and is drawn in the figure as the y-axis. This discriminant function is the theoretical function that assumes the means and covariances are exactly known. In practice, these 200 points are needed to construct estimates $\bar{x}$ for the means and S_N for the covariance for each of the classes. When this is the case, the resulting discriminant function is near the line $x = 0$, but for these particular data the discriminant function is drawn in figure 2, along with the theoretical estimate. Figure 2a gives a local view of the discriminant and figure 2b gives more of a global view of the discriminant function; points to the left of the boundary are assigned to class 2 and points to the right of the boundary to class 1.

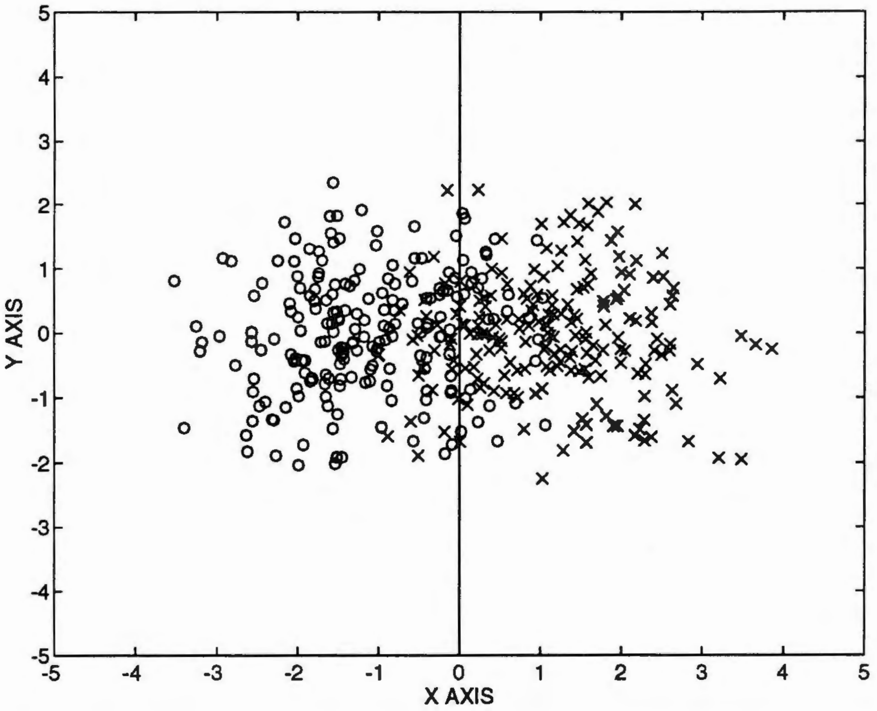
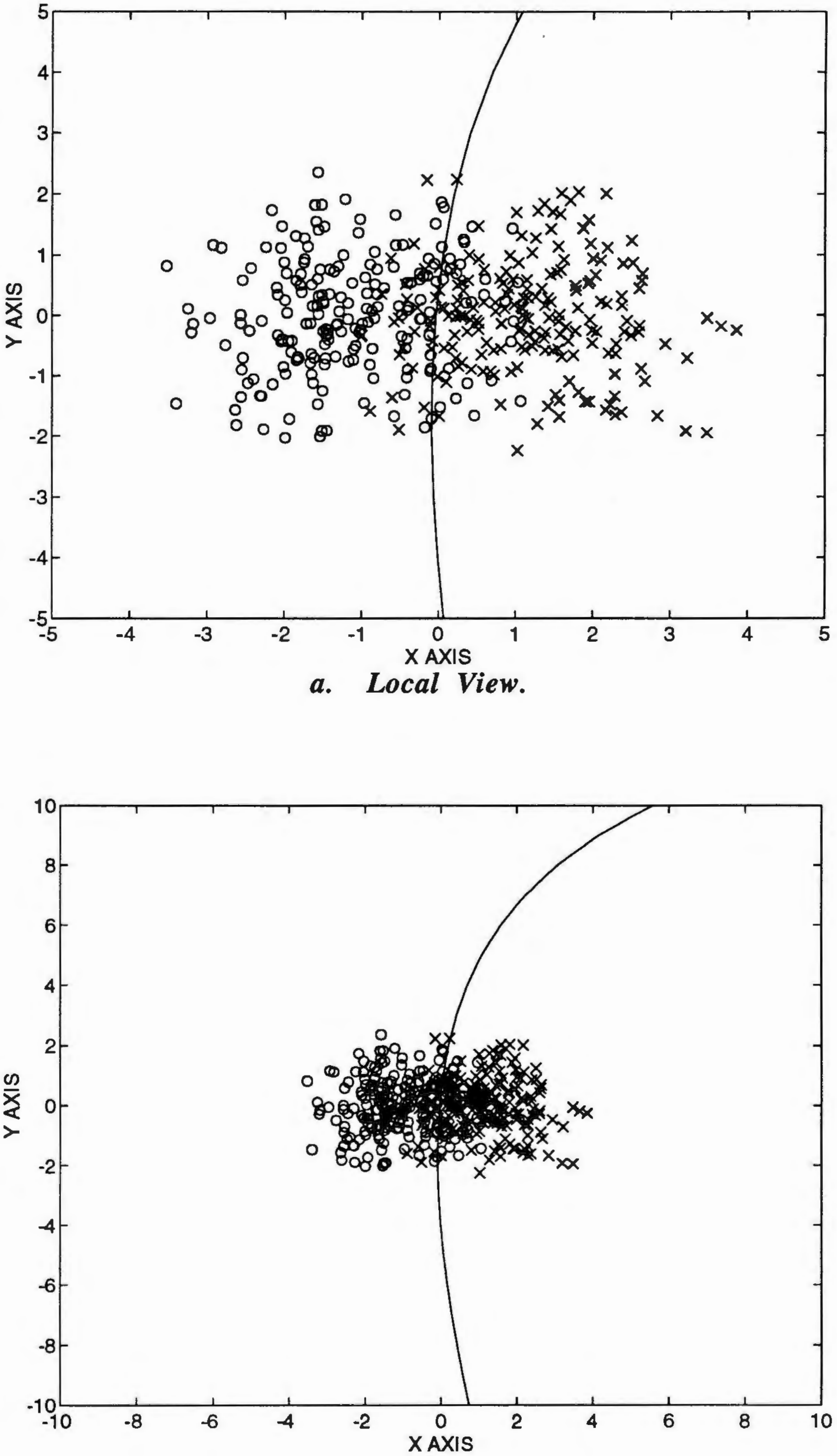


Figure 1. Scatter Plots of the Normal Training Data.

The covariance matrices for both the ideal and the sample covariance are shown in table 1. The sample covariance is close to the ideal covariance estimate as one expects. So, in summary, one sees that the parametric representations work well provided the assumptions used to derive them are valid. For this data sample, the ideal centering and covariance estimate yields 13.5 percent error, which is close to the theoretical 10.0 percent. The sample statistics yield 12.75 percent error. And in section 5, it will be shown that the robust parameter estimates yield 12.25 percent error. For this type of data, one expects all the methods to work well and all to be close to the theoretical results. Moreover, to test more thoroughly, one would need more extensive simulation results. In the next section, it will be demonstrated that when the distribution assumptions are not valid, the results quickly degenerate.



a. Local View.

b. Global View.

Figure 2. QC Generated by Sample Statistics Using Normal Training Data.

Table 1. QC Parameter Estimates, Known vs Sample Statistics, Normal Data

Ideal mean vector for the class 2 cluster.					Sample mean vector for the class 2 cluster.				
-1.27	0.0	0.0	0.0	0.0	-1.161	0.007	-0.016	0.005	0.068
Ideal mean vector for the class 1 cluster.					Sample mean vector for the class 1 cluster.				
+1.27	0.0	0.0	0.0	0.0	1.150	-0.027	0.084	-0.015	0.047
Ideal covariance matrix for the class 2 cluster.					Sample covariance matrix for the class 2 cluster.				
1.0	0.0	0.0	0.0	0.0	0.963	0.101	0.125	0.035	-0.007
0.0	1.0	0.0	0.0	0.0	0.101	0.898	0.004	0.003	-0.002
0.0	0.0	1.0	0.0	0.0	0.125	0.004	1.097	0.036	0.067
0.0	0.0	0.0	1.0	0.0	0.035	0.003	0.036	0.926	-0.003
0.0	0.0	0.0	0.0	1.0	-0.007	-0.002	0.067	-0.003	0.957
Ideal covariance matrix for the class 1 cluster.					Sample covariance matrix for the class 1 cluster.				
1.0	0.0	0.0	0.0	0.0	0.987	-0.028	0.013	0.160	-0.162
0.0	1.0	0.0	0.0	0.0	-0.028	0.834	0.041	-0.007	0.107
0.0	0.0	1.0	0.0	0.0	0.013	0.041	0.859	-0.075	-0.052
0.0	0.0	0.0	1.0	0.0	0.160	-0.007	-0.075	1.047	-0.166
0.0	0.0	0.0	0.0	1.0	-0.162	0.107	-0.052	-0.166	1.030

3. EFFECT OF NON-NORMAL DATA ON CLASSIFIER DESIGN

When the data are normal, the QC as expected exhibits good performance. However, anomalous behavior occurs when there are outliers in the data. Outliers and non-normal data are problem areas for classifier design. In this section both of these effects are discussed. First, the outliers are modelled as a cluster of mistaken data and their effect on the sample statistics is not only analyzed, but also illustrated with an example. Second, the non-normal data are represented as Cauchy data and then their effect on the sample statistics and the classifier design is illustrated.

3.1 EFFECT OF OUTLIERS

Outliers are data samples that do not conform to our concept of the distribution model. Their cause may be either inadequate modeling or errors in the data or both In either case, one would like to minimize the effects of these outliers in training the classifiers without manually inspecting all the data. Moreover, even if the data are inspected manually, in high dimensional cases, visual methods may be inadequate and clustering may be needed to spot these outliers. The abnormal influence that outliers produce on the training is due primarily to the linearity of the sample statistics used to estimate the first and second order parameters in the classifier. One way around this is to use resistant estimation techniques, which implicitly censor the data by reducing the influence of outliers on the estimate. This will be discussed in detail in section 4.2.

To demonstrate the effect of outliers on the QC, the following two-class pattern recognition problem is analyzed. Assume one has two normal clusters with unit variance. The first cluster is distributed $N(\boldsymbol{m}_1,\Sigma_1)$ and the second is distributed $N(\boldsymbol{m}_2,I)$ where $\boldsymbol{m}_1 = [m_b,0,0,0,0]$, $\boldsymbol{m}_2 = [-m,0,0,0,0]$, and $m_b = (1-\varepsilon)m + \varepsilon b$. Here ε is the probability that a sample in class 1 will be an outlier. For $\varepsilon = 0$, this problem is the standard two-class antipodal signal problem. The solution is known. For equiprobable classes, the optimal Bayesian decision rule for a sample vector $\boldsymbol{x}^t = [x_1,x_2,x_3,x_4,x_5]$ is simply: decide class 2 if $x_1 > 0$; otherwise, decide class 1.

Now suppose a small percentage of the outliers in the class 1 sample are outliers centered about the point (b,0). This is the contaminated normal model. If the proportion of outliers is ε, then the covariance matrix of class 1 is not I, but instead,

$$\Sigma_1 = E(\boldsymbol{x}-\boldsymbol{m}_1)(\boldsymbol{x}-\boldsymbol{m}_1)^t =$$

$$(1-\varepsilon)\Big[E(\boldsymbol{x}-\boldsymbol{m}_1)(\boldsymbol{x}-\boldsymbol{m}_1)^t|E\boldsymbol{x}=\boldsymbol{m}_b\Big] + \varepsilon\Big[E(\boldsymbol{x}-\boldsymbol{m}_1)(\boldsymbol{x}-\boldsymbol{m}_1)^t|E\boldsymbol{x}=\boldsymbol{b}\Big],$$

which yields $\Sigma_1 = I + \Delta\boldsymbol{m}\Delta\boldsymbol{m}^t$ where $\Delta\boldsymbol{m}^t = \sqrt{\varepsilon(1-\varepsilon)}(\mathrm{b}-\mathrm{m})[1,0,0,0,0]$.

To see how the solution to the two-class problem is altered, consider the following formulation of the two-class problem. For the QC (reference 1, p.54), the decision statistic and critical region is given by

$$h(\boldsymbol{x}) > \frac{1}{2}\log\frac{P(\omega_1)}{P(\omega_2)},$$

where

$$h(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_1)^t\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{m}_1) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_2)^t\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{m}_2) + \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|},$$

and the means and covariance matrices are given above. The inverse for Σ_2 is I. The special form of Σ_1 produces a convenient representation for the inverse given by reference 1, p.43. If $S = \Sigma + \boldsymbol{bb}^t$, then its inverse is given by $S^{-1} = \Sigma^{-1} - \frac{\Sigma^{-1}\boldsymbol{bb}^t\Sigma^{-1}}{\mathit{1} + \boldsymbol{b}^t\Sigma^{-1}\boldsymbol{b}}$ where $\Sigma = \mathit{I}$. So then, applying this formula to Σ_1 yields $\Sigma_1^{-1} = \mathit{I} - \frac{\Delta \boldsymbol{m} \Delta \boldsymbol{m}^t}{\mathit{1} + \varepsilon(1-\varepsilon)(\mathrm{b} - \mathrm{m})}$, which simplifies the quadratic discriminant. The above inverse can be written in the following form: $\Sigma_1^{-1} = I - C$ where $C = diag\left[\frac{c}{1+c},0,0,0,0\right]$ and $c = \varepsilon(1-\varepsilon)(m-b)^2$. Using this form of the matrix inverse in the discriminant yields

$$h(\boldsymbol{x}) = \boldsymbol{x}^t(\boldsymbol{m}_2 - \boldsymbol{m}_1) + \frac{1}{2}(\boldsymbol{m}_1^t\boldsymbol{m}_1 - \boldsymbol{m}_2^t\boldsymbol{m}_2) + \frac{1}{2}\ln(1+c) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_1)^t C(\boldsymbol{x} - \boldsymbol{m}_1),$$

where the first two terms represent the unperturbed two class I vs I decision rule and last two terms represent the impact of the outliers. Setting h(x) = 0 gives the following equation for the decision boundaries:

$$(x - m_c)^2 = m_c^2 - m_b^2 + \frac{1}{2}(m_b^2 - m^2)r + \frac{r}{2}\ln(1+c) \ ,$$

where

$$m_c = m_b - \frac{r}{2}(m + m_b) \text{ and } r = \frac{2(1+c)}{c} \ .$$

The equation represents a pair of vertical lines centered about m_c. The decision rule is: decide class 2 if x_1 is between the vertical lines, else decide class 1. Figure 1 illustrates the decision region for a two-class I vs I problem with no outliers and figure 3 illustrates the decision region where there are outliers near (b,0). The clusters were generated by a normal random number generator and plotted using Matrix Laboratory (MATLAB). In this example, m=1.27, $\varepsilon = 0.1$, and $b = -20$, which implies that 10% of the points of class 1 are outliers located in a normal cluster with mean of -20. This solution represents the QC when there is no error in estimating the means and covariances for the two classes.

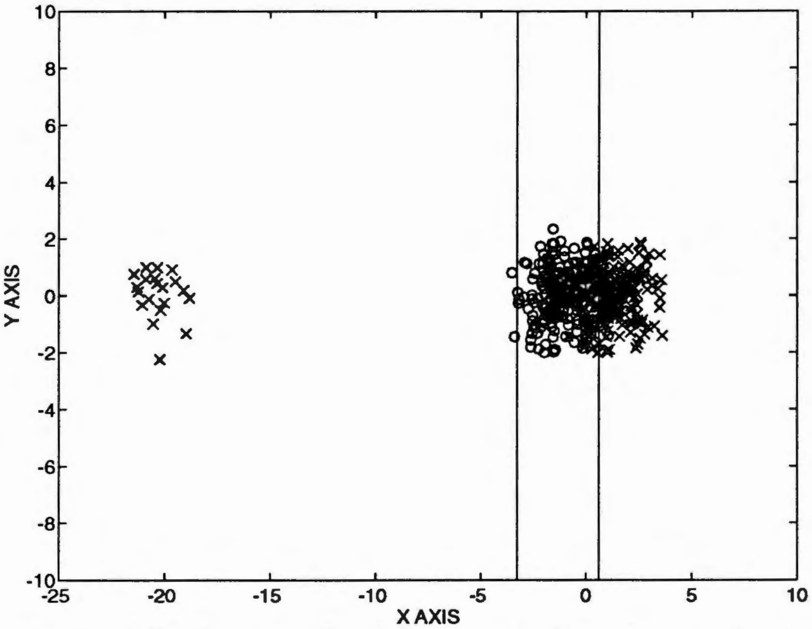
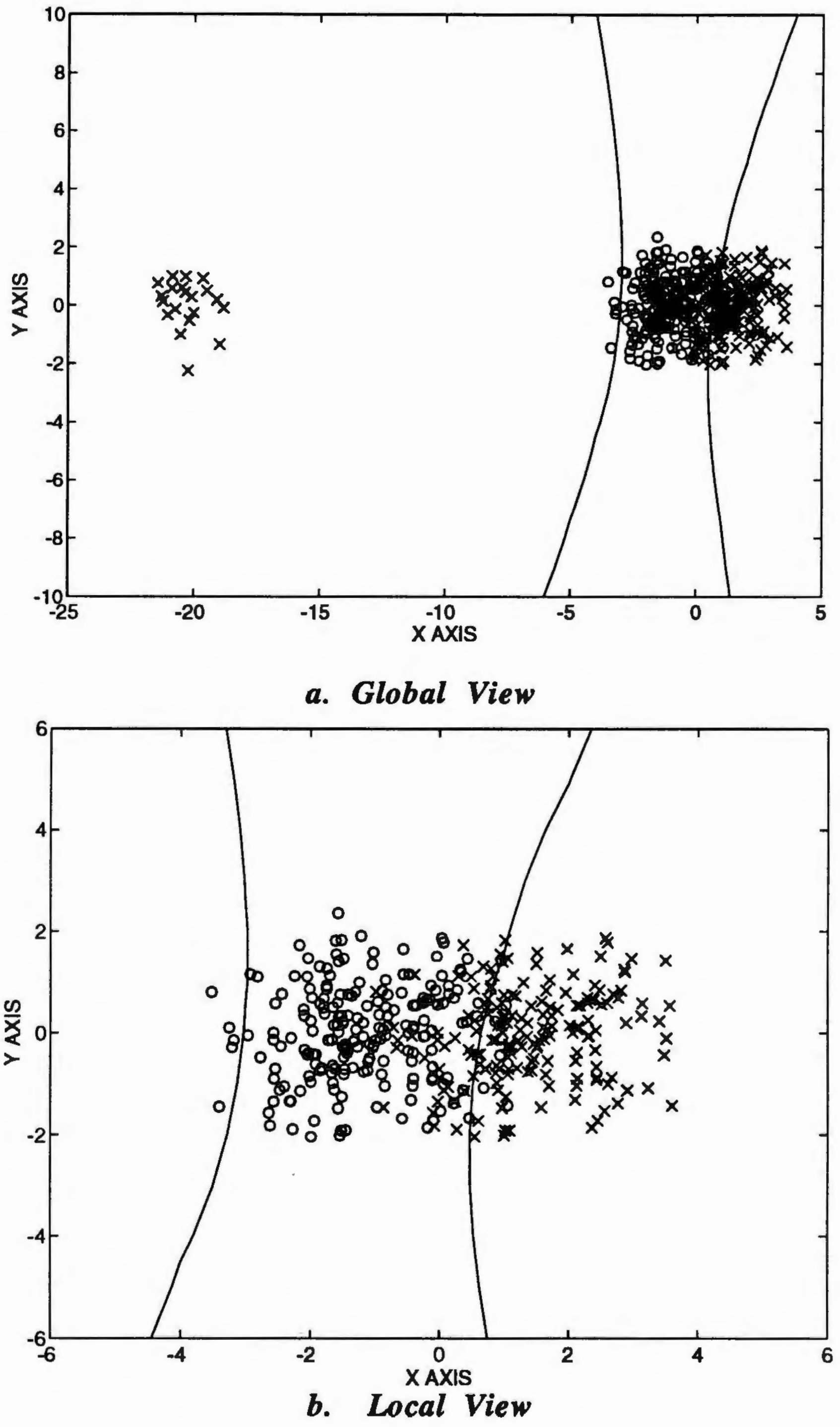


Figure 3. QC Generated with Known Population Statistics

The estimation of the means and covariance matrices of the two classes introduces quite a bit of error, in part because the samples are normal, but also because of the outliers introduced into the class 1 samples. The first and second order statistics are distorted because of the linearity of the sample mean and the sample covariance. The sample moments and the true sample moments are given in table 2. Figure 4 shows the decision regions of the QC for the sample moments; figure 4a is the global view and figure 4b is the local view. Note that the quadratic and linear terms other than x_1 no longer cancel out exactly, which gives rise to the hyperquadrics (reference 2, p. 30). In the first two dimensions in figure 4, a hyperbola is shown. The effect of the outliers is to push the RHS boundary into the class 1 cluster causing more errors, and the LHS boundary excludes the outliers in the class 1 data, but not as effectively as the ideal classifier illustrated in figure 3.

Table 2. QC Parameter Estimates, Known vs Sample Statistics, Contaminated Normal Data

Ideal mean vector for the class 2 cluster.					Sample mean vector for the class 2 cluster.				
-1.27	0.0	0.0	0.0	0.0	-1.161	0.007	-0.016	0.005	0.068
Ideal mean vector for the class 1 cluster.					Sample mean vector for the class 1 cluster.				
+1.27	0.0	0.0	0.0	0.0	-0.783	-0.038	0.047	-0.006	0.056
Ideal covariance matrix for the class 2 cluster.					Sample covariance for the class 2 cluster.				
1.0	0.0	0.0	0.0	0.0	0.963	0.101	0.125	0.035	-0.007
0.0	1.0	0.0	0.0	0.0	0.101	0.898	0.004	0.003	-0.002
0.0	0.0	1.0	0.0	0.0	0.125	0.004	1.097	0.036	0.067
0.0	0.0	0.0	1.0	0.0	0.035	0.003	0.036	0.926	-0.003
0.0	0.0	0.0	0.0	1.0	-0.007	-0.002	0.067	-0.003	0.957
Ideal covariance matrix for the class 1 cluster.					Sample covariance for the class 1 cluster.				
2.9	0.0	0.0	0.0	0.0	43.445	-0.049	-0.013	0.307	0.461
0.0	1.0	0.0	0.0	0.0	-0.049	0.814	0.018	-0.039	-0.016
0.0	0.0	1.0	0.0	0.0	-0.013	0.018	1.094	0.004	0.084
0.0	0.0	0.0	1.0	0.0	0.307	-0.039	0.004	0.907	0.041
0.0	0.0	0.0	0.0	1.0	0.461	-0.016	0.084	0.041	0.904



a. Global View

b. Local View

Figure 4. QC Generated by Sample Statistics Using Contaminated Normal Data.

In section 4, it will be illustrated that the robustified version of the quadratic classifier does far better than the parametric versions considered in this section. In addition, the probability of error will improve. At this point, it is clear that the outliers have deteriorated the performance of the quadratic classifier.

3.2 EFFECT OF NON-NORMAL DATA

Normal data are very well behaved in the sense that their distribution tails $P(X > x)$ for large x are exponentially small. These light tails guarantee that few samples are found more than 5σ from the mean. However, in heavy-tailed distributions like the exponential distribution or the Cauchy density, random variates give the impression that they are outliers compared to normal samples. Thus, the linear statistics, which are optimal for the normal distribution, perform poorly for other distributions. Since the underlying noise distribution is probably never known, the linear and quadratic discriminants must be robustified by using robust estimates of the parameters just as in the outlier case.

As an example, consider Cauchy samples that have been normalized by twice the MAD of the samples. The samples are shown in figure 5. If sample statistics are used to construct the quadratic classifier, the results are catastrophic. The optimum discriminant using the maximum likelihood detector is

$$h(x) = -\ln\left[\frac{1+\left(\dfrac{x-m_2}{S}\right)^2}{1+\left(\dfrac{x-m_1}{S}\right)^2}\right] - t \ ,$$

where t is the threshold. Similar to the normal sample, the y-axis forms the decision boundary; that is, if m_2 is -m and m_1 is m, then t=0 is the threshold for h(x) so that one decides class 1 if h(x) < 0 and class 2 otherwise. Using sample statistics to estimate the mean and covariance produces a quadratic decision region as shown in figure 6, where the boundary forms a hyperbola and yields 48.75% error. Using the median for the centering constant and the robust estimate for the covariance matrix also produces a quadratic discriminant with a hyperbolic boundary, but it has much lower error. This is discussed in detail in section 5.

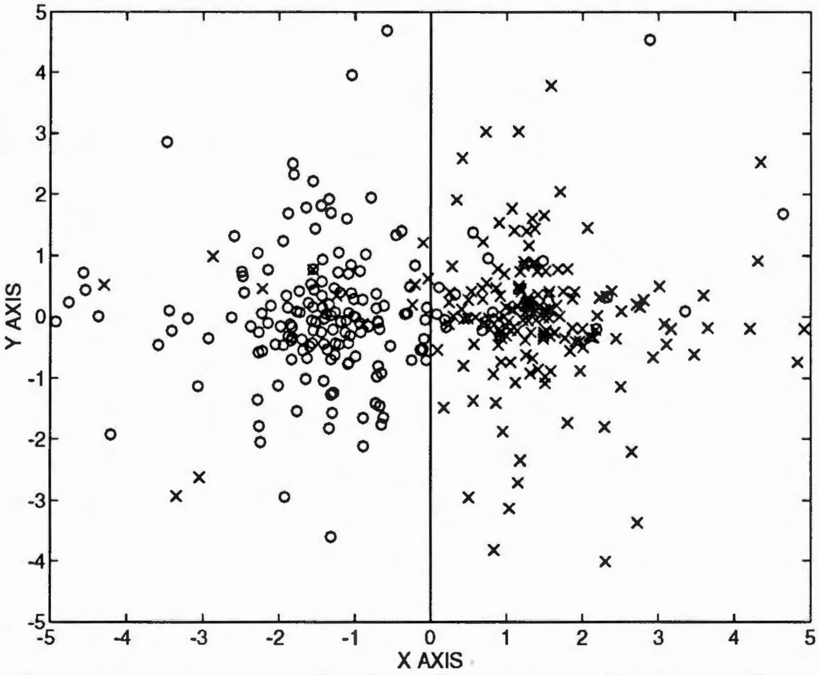


Figure 5. Scatter Plot for the Cauchy Training Data

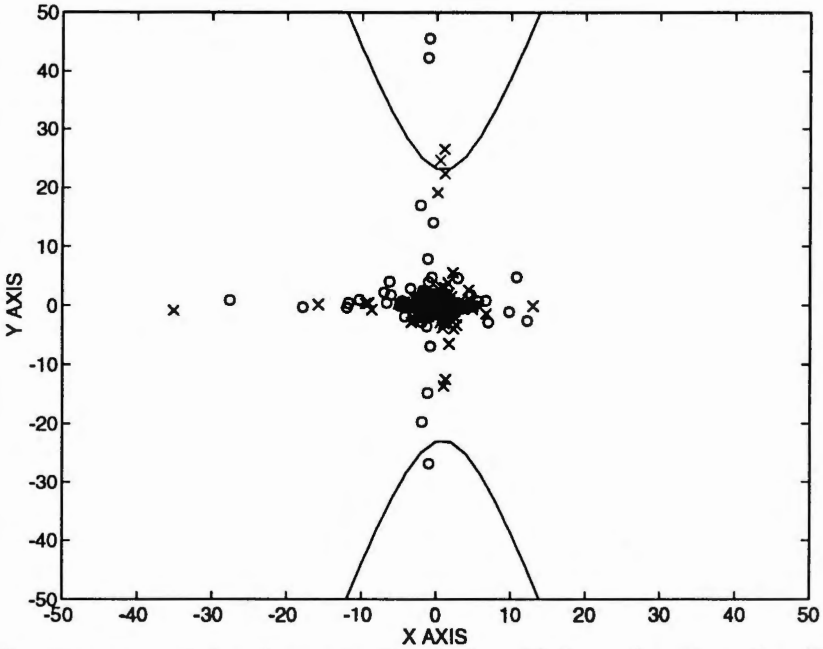


Figure 6. QC Generated by Sample Statistics Using the Cauchy Training Data

Ideally, one would not use a QC on Cauchy data. Practically, one might prefer to standardize the QC and use robust parametric estimates so that it works effectively across a large class of unimodal density functions. The robust statistics seem to be resistant to even the Cauchy data, which is commonly used as the worst case example in nonparametric statistics.

4. ROBUST STATISTICS

In this section, robust statistics are applied to the estimation of the parameters of the QC. This means the sample mean is replaced by a median vector and the sample covariance is replaced by a robust covariance estimate. First, robust statistics are discussed to give a feeling for the reasons why this approach is used. Second, some robust statistics are illustrated that apply to the estimation of the QC parameters. This section discusses only one-dimensional statistics to make the visualization easier to grasp. Third, the robust estimation of the covariance matrix is discussed. This algorithm is a recursive procedure requiring a good initial estimate. Fourth, the initialization of the covariance estimate is discussed, using three different techniques. Examples of this technique are in the next section.

4.1 BACKGROUND

In robust statistics, there are different types of robustness (reference 3, p. 40-47). The first is called *qualitative robustness*, a necessary condition that describes the response of the statistic with respect to small perturbations or "wiggling" of the data. The second quantifies the effects of perturbations of the underlying distribution on the statistic. The quantitative information is provided by the *influence function*, which is a functional derivative of the statistic with respect to a change in one of the data points or the data itself. The third type of robustness is the *breakdown point*. Intuitively, the breakdown point is the minimum percentage of outliers in the data needed to destroy the reliability of the estimator. Only the last type of robustness is the most important for this memorandum. For a complete discussion of robust statistics there are several texts the reader might consult (references 3-7). This discussion is only cursory, attempting to give the reader a cursory overview.

Robust statistics are *not* the same as nonparametric statistics. In nonparametric statistics, one designs statistics that do not assume any particular form of a parametric distribution, or if a class of distribution functions is specified, do not assume a distributional form. An example of a nonparametric statistic is the sign test. Here, one uses the statistic

$$Sgn = \sum_{i=1}^{N} sgn(X_i - c),$$

where the sum of the individual signum functions for each sample has thrown away all the magnitude information. Defining the class of distributions as the set of all symmetric distributions, one can apply the statistic Sgn to detect change in location. In this case the null hypothesis is $c = 0$, and the alternative hypothesis is $c \neq 0$. A change in location of the sample to the right will cause the sign test to be more positive. The power and the level of the statistical hypothesis test is independent of the form of the underlying distribution. In fact, the power function and the level are described by the binomial distribution B(p,n), where p is the $P(X_i - c > 0)$ and c is the shift in location (reference 8, p. 103).

Robust statistics are *not* the same as parametric statistics. Parametric statistics are based upon a specific distributional form, whereas robust statistics admit not only the parametric forms but also distributions "near by." In figure 7, the squares represent the class of all distribution functions, which in general has infinite dimensions. Practical parametric distributions are specified on finite dimensional Euclidean spaces. These figures are similar to those used by Hampel (reference 3, p.7,10). In figure 7a the shaded areas represent the class of distributions on which the nonparametric distributions apply. This is a non-trivial portion of the set of distributions. The set of parametric distributions is shown in figure 7b and is a tiny portion of all distributions, since it is specified on finite dimensional space and has a finite number of parameters. Figure 7c is the set of all robust distributions, which is a superset of the parametric distributions. Figure 7d is the robust set of distributions, that surround the multivariate normal distribution and represents the set of distributions needed to robustify the quadratic discriminant. Note that here only one specific parametric form is the center of the class, and the set of distributions for which the robust statistic is designed is a "neighborhood" of the specific form. Therefore, Hegal's philosophy has prevailed; the robust statistics form a class of statistics between the two "competing philosophies" of parametric and nonparametric statistics.

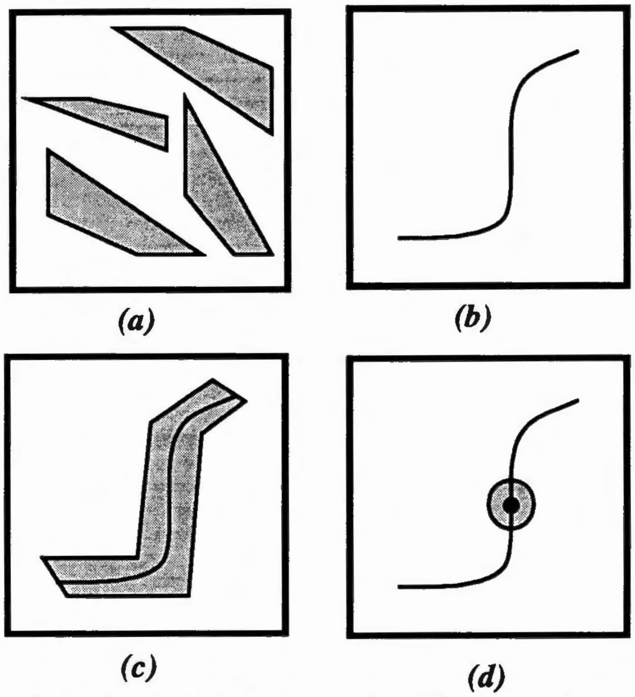


Figure 7. Probability Distributions for Various Types of Statistics.

So far, all that has been said is what robust statistics is not. What is it then? The following working definition is taken from Hampel (reference 3, p. 7).

> "Robust statistics, as a collection of related theories, is the statistics of approximate parametric models. It is thus an extension of classical parametric statistics, taking into account that parametric models are only approximations to reality."

Robust statistics include the concept of outlier rejection, but are not defined by the concept. Some claim that robust statistics is a natural extension of outlier rejection (reference 3, p. 71) and, certainly, outlier rejection is an important consequence of using robust statistics. Parametric statistics are derived assuming that all the samples are random variates from a distribution form with a fixed set of parameters. Robust statistics assume that the majority of the samples come from a particular distributional form with fixed parameters and the rest of the samples are outliers, or samples from a different distribution closely related to the first distribution. It is this particular aspect of the robust statistics that is used to construct robust estimates of the parameters in the QC.

To study the various types of robustness Hampel developed a special tool. The Influence Function (IF) is a mathematical tool that quantifies the effects of outliers and is also a heuristic tool that describes the phenomena seen in robust statistics (reference 3, p. 83). Formally, the influence function is defined as (reference 3, p. 84):

$$IF(x;T,F) = \lim_{t \downarrow 0} \frac{T((1-t)F + t\Delta_x) - T(F)}{t},$$

where F is a cumulative distribution function (CDF) in the set of all distributions defined on the sample space χ, and Δ_x is the CDF that puts probability mass 1 at the point x. One might call this the degenerate distribution at point x. The limiting operation is a Gateaux differential of the functional T. Moreover, the $IF(\cdot)$ also can be written in the more familiar functional form of an integral (reference 3, p. 83) :

$$IF(x;T,\Delta_x) = \int a(x)d\Delta_x = \int a(u)\delta(u-x)du,$$

where $a(x)$ acts as the impulse response function of the impulse placed at the point x. Hence, the IF measures the "response" of the statistic to a sample point inserted at the value x.

The power of this representation is that the nonlinear instantaneous transformation $IF(\cdot)$ filters the samples, thus producing a new set of samples that converges to the normal distribution. To see this, one has to first realize that any statistics can be written as a functional on the distribution function, i.e., $T_N = T_N(X_1,\cdots,X_N) = T(F_N)$. For example, the sample mean is written as $\overline{X}_N = \int x dF_N$, where F_N is the empirical cumulative distribution function (ECDF) for the sample. Assume that the ECDF converges to a CDF F with probability one by the Glivenko-Cantelli theorem, i.e., $\lim_{N \to \infty} F_N \to F \; a.e.$, where the abbreviation a.e. means almost everywhere. Then,

$$\sqrt{N}(T_N - T(F)) \approx \frac{1}{\sqrt{N}} \sum_{i=1}^{N} IF(X_i;T,F) + Higher\ Order\ Terms\ ,$$

and $\sqrt{N}(T_N - T(F)) \to N(0,V(T,F))$ in law by the Central Limit Theorem. Therefore, the sum of the samples filtered through the $IF(\cdot)$ is asymptotically normal with a specified variance given by $V(T,F) = \int IF(x;T,F)^2 dF(x)$. The variance then looks like the samples have been merely filtered through the IF. Intuitively, $IF(\cdot)$ is acting as a limiting function that truncates the random variables.

The robustness of a statistic $T(F)$ can be explained by the properties of the IF. For example, the *gross error sensitivity* of $T(F)$ at the distribution function F is defined as

$$\gamma^*(T,F) = \sup_x |IF(x;T,F)|,$$

where x is over the domain where the distribution is defined or the probability space defined by F. This error function upper bounds the damage any single sample can do to the estimator. So if the IFs as defined are bounded, then it is much easier to satisfy the conditions for asymptotic normality. If $\gamma^* < \infty$ the statistic is referred to as B-robust.

This quantitative tool can also be used to define *local-shift sensitivity* or, more intuitively, the worst case response from taking the data set sample point and "wiggling" it. The worst local derivative is defined as

$$\lambda^* = \sup_{x \neq y} \frac{|IF(y;T,F) - IF(x;T,F)|}{|y-x|},$$

which may be infinite for finite jumps in the IF. An example of this occurs when T is the median and its IF is the sign function. Because of the discontinuity about $y = x = 0$, this value is infinite. So the median, which is always thought of as a robust estimator of center location, is sensitive to local shifts. However, the statistic is B-robust.

The next measure of robustness is the breakdown point, which is roughly defined to be the percentage of outliers needed to render the statistic useless. The sequence of estimators $\{T_n\}_{n=1}^{\infty}$ converges to a point; i.e. it is a point estimator, which assumes the samples are independent and identically distributed random variables from the distribution F. However, if enough samples come from another distribution, say the distribution G, then the point estimator becomes nonsense. That is, it may not converge or, if it does converge, it may not converge to anything that is remotely near the point estimate when all the samples are from F. If G is "very far" from F, then fewer samples from G are required to break down the estimator performance. To quantify this, one needs to define a metric between distribution functions; here one can use either the Prohorov metric or the Levy metric. Hampel uses the Prohorov metric, $\pi(F,G)$ (reference 3, p. 96) . Breakdown is defined then by looking at the behavior of the sequence of estimators (reference 3, p. 97). By finding the greatest distance between F and G such that the sequence $\{T_n\}_{n=1}^{\infty}$ still converges, one has found the maximum resistance to variation of the assumptions.

Mathematically, the breakdown point ε^* for a sequence of estimators is defined at F as (reference 3, p. 97)

$$\varepsilon^* = \sup_{\varepsilon \leq 1} \{\exists \, compact \; set \; K_\varepsilon \subset \Theta \; st \; \pi(F,G) < \varepsilon \Rightarrow G(\{T_N \in K_\varepsilon\}) \to 1 \; as \; N \to \infty\},$$

where Θ is the parameter space. Intuitively, this definition says: if F is close enough to G, then the sequence of estimators converges to a point in the parameter space or to a very small region around the true value. However, if F is not close enough to G, then the convergence breaks down, which means there is no small region about the true value in which the estimate is found with high probability. If the compact region were a small interval about the true parameter value, the requirement would be much like the definition of convergence in probability.

If one uses another type of metric, where $\pi(F,G) < \varepsilon$ means $G \in \{(1-\varepsilon)F + \varepsilon H\}$, the ε represents the percentage of the data having the distribution H that is mixed into the data set. This is the genesis of the intuitive interpretation of breakdown as the percentage of outliers required to make the estimator nonsensical. The ε breakdown measures the global resistance of the estimator to contaminated samples, and thus is a key parameter. In addition, to these quantitative measures, one has other qualitative measures that Hampel has also defined (reference 3, p. 98). Both the median and the MAD estimators are excellent examples of robust statistics and these are discussed in the next section.

4.2 ONE-DIMENSIONAL ROBUST STATISTICS

In one dimension, the median and median absolute deviation (MAD) from the median estimators are robust statistics. If the data sample is denoted as $X = \{X_1, X_2, \cdots, X_N\}$, then the median of this sample is denoted as $med(X)$. For an ordered sample $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$, the median is defined as $\left[X_{(k+1)} + X_{(k)}\right]/2$, $N = 2k$, if N is even and as $X_{(k+1)}$, $N = 2k+1$, if N is odd (reference 4, p. 11). The median is resistant to outliers since its breakdown point is 1/2. Again, the breakdown point of a statistic is loosely defined as the fraction of outliers at which the statistic no longer is a reliable estimate. Likewise, the MAD is a measure of dispersion, defined by $S_N = med_i\{|X_i - med(X)|\}/0.6745$ and has a breakdown point of 1/2 (reference 5, p. 107). Note the scale constant 0.6745 has been introduced so that if the samples $\{X_i\}_{i=1}^{N}$ are standard normal, then $S_N \to 1, as\ N \to \infty$ (reference 4, p. 237). Both the median and the MAD are resistant to outliers since their influence functions are bounded and large variations of the data can produce only bounded variations of the statistic. However, the local-shift sensitivity for the median is infinite. This is caused by the fact that the influence function for the median is the signum function, i.e., $sgn(x)$, and variations of the data points about the value $x = 0$ can produce jumps in the influence functions leading to infinite derivatives.

In contrast, the IF for the sample mean is the identity function $\psi(x) = x$, and for the variance is $\psi(x) = x^2 - 1$, which is continuous so small variations of the sample values produce correspondingly small variations of the IF. This causes the local-shift sensitivity to be minimal for the mean (reference 5, p. 22). However, the sample mean and variance are not resistant to outliers since their influence functions are unbounded, implying the gross error sensitivity is infinite. In this memo, the gross-error breakdown point is of primary concern since it is a global measure of the resistance to outliers. So except for the time and space complexity, the median and MAD are a better choice in uncertain data environments.

For this memo, the vector median is defined as the vector of medians. (This is not the only median that can be used.) With the median vector as the centering constant, the robust version of the covariance matrix is based on Huber's M-estimate applied to the covariance (reference 9). In this algorithm, the average of the outer product of the sample vectors or the sample covariance is replaced by a weighted average of outer products. The weighted average is obtained recursively, where each weight is a non-increasing function of the distance from the center of the whitened data. This method may be viewed as a continuous version of outlier rejection based on the distance the outliers are from the median in terms of standard deviations. The algorithm is explained in detail in Huber's text (reference 5, p. 238), and the algorithm presented in this memorandum is only a slight modification of his algorithm in which the location estimation step is ignored, since the centroid is assumed to be the median vector.

One issue of importance is the breakdown of this covariance estimation technique. Maronna (reference 9) and Huber (reference 5) point out that the breakdown of the procedure is on the order of 1/d where d is the dimension of the sample space. For example, if the dimension of the data is five as in the example used in this memo, a worst case analysis shows that only 20% of the data as outliers can produce breakdown of the statistic. For higher dimensional spaces, this breakdown is small since if d=100, only 1% of the samples can cause breakdown. This has not been observed in practice as yet but that may only be a matter of insufficient exercise of the algorithm. Huber also suggests a possible way to detect this breakdown and reject outliers to repair it (reference 5, p.

228). Other alternatives to increase the breakdown are currently under study. In any case, it should be remembered that using the sample covariance means that only one sample is needed to break down that estimate, so although 1/d is a disappointing breakdown point, it is certainly better than the sample covariance estimate.

Three one-dimensional examples are given to illustrate the resistance of robust statistics to heavy-tailed distributions and outliers. The first example uses samples from a normal distribution, the second from a contaminated normal distribution, and the third from a Cauchy distribution. Figure 8a gives an example for 200 samples along the x-axis from a N(1.27,1) sample. The true mean is marked by a '+' for origin, the sample mean is marked by an 'o' and the median is marked by an 'x'. Figure 8a only has an 'o' and an 'x' to indicate the median and the mean, respectively; here, they are so close to each other that the marks, located near (1.27,78) appear to coincide. Note that the median will be one of the sample values if the sample size is odd; otherwise, it will be the average of two samples. For the normal distribution, the sample mean, the mode, and the median should be near 1.27 so that one expects, even for sample sizes of 200, that these statistics will coincide. For heavier-tailed distributions like the Cauchy, one cannot be sure these three statistics will coincide, even for large samples. For the contaminated normal distribution where the samples come from $N(1.27,1)$ with probability $1-\varepsilon$ and from some distribution centered at -20 with probability ε, these three statistics need not coincide. If one views the samples located at -20 as outliers, figure 8b illustrates how the zero point, the mean, and the median are related. This figure demonstrates the vulnerability of the sample mean to outliers. The sample mean, shown as 'o' near the top of the figure, is pulled away from the true mean of 1.27 for the uncontaminated sample, but the median, denoted by 'x', is still very close to the true mean. Figure 8c illustrates a Cauchy sample of the 200 data points that behaves well and centered at -1.27, and figure 8d shows a Cauchy sample centered at +1.27 that does not behave well. In the latter case, a few extremely large outliers destroy the mean estimate completely, yet the median still stays close to the true center of the data cluster, which is indicated by a '+' symbol.

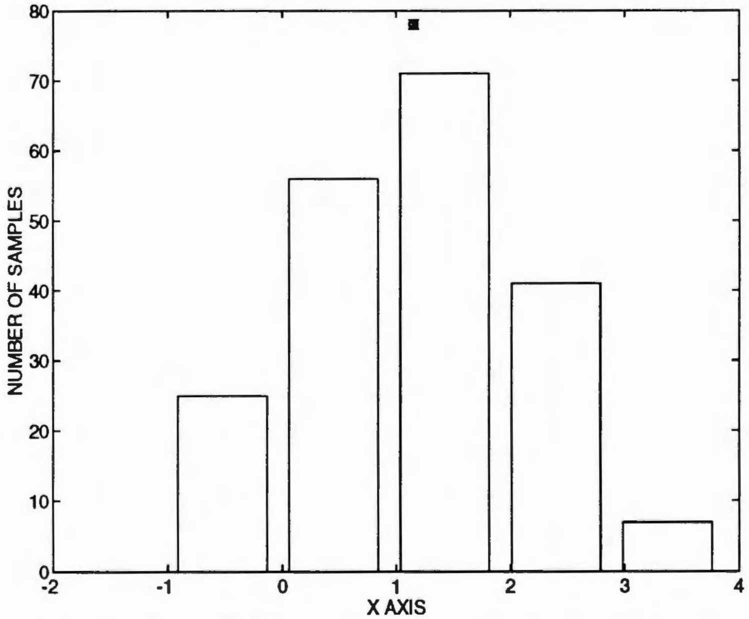


Figure 8a. Histogram of a $N(1.27,1)$ *Sample*
(The mean and median are shown on the top of the plot.)

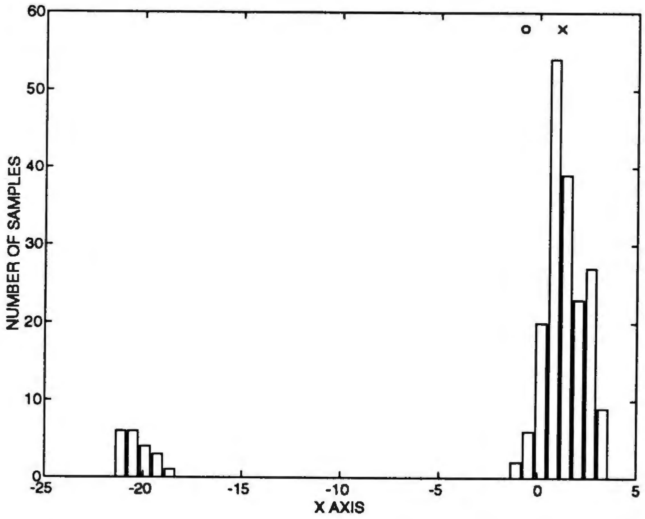


Figure 8b. Histogram of a $N(1.27,1)$ *Sample Contaminated by Outliers (The location of the mean, median, and true cluster center shown at top.)*

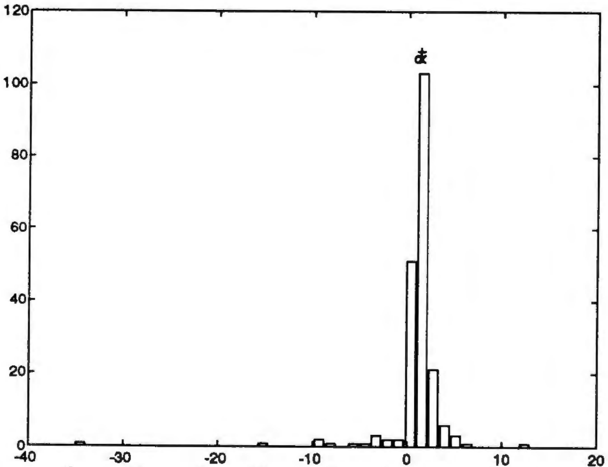


Figure 8c. Histogram of a Cauchy Sample Centered at +1.27 Along the X-Axis

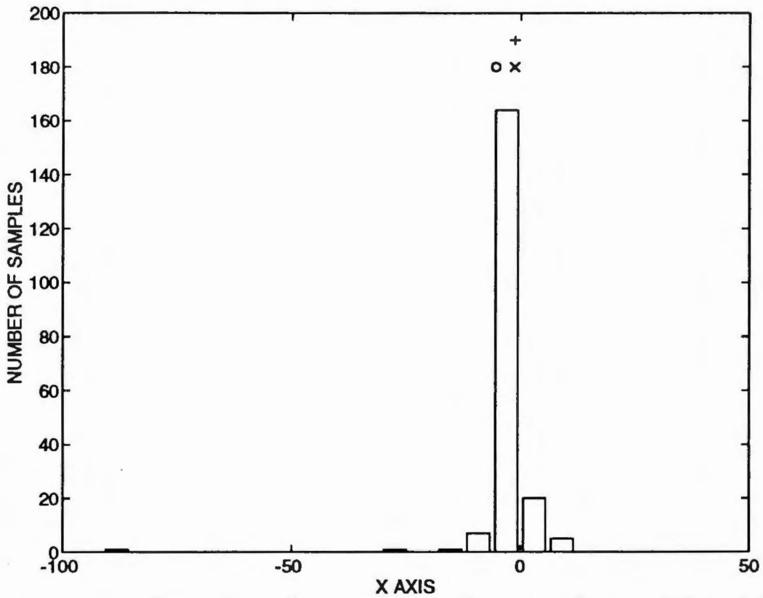


Figure 8d. Histogram of a Cauchy Sample Centered at -1.27 Along the X-Axis

The next set of examples illustrates dispersion estimates. On the same data sets used in the above examples, the sample covariance and the MAD estimate are illustrated. In figure 9a, the normal sample illustrates the true two standard deviation (STD) window, the sample STD estimate derived from the sample covariance and the window resulting from the scaled MAD estimate. Not surprisingly, for the normal sample these estimates are fairly close. In figure 9b, the same statistics are shown using a sample from the contaminated normal distribution. Notice how large the sample STD is, yielding unrealistically higher STD. The same estimates are obtained for the ill-behaved Cauchy distribution and are illustrated in figure 9c. Notice the stability of the scaled MAD estimate throughout the examples and the instability of the sample. Again, '+' represents true STD about the true mean, when that information exists. For figure 9c, only the center value is given since, theoretically, moments do not exist for Cauchy data. Note that sample STD has broken down as a statistic.

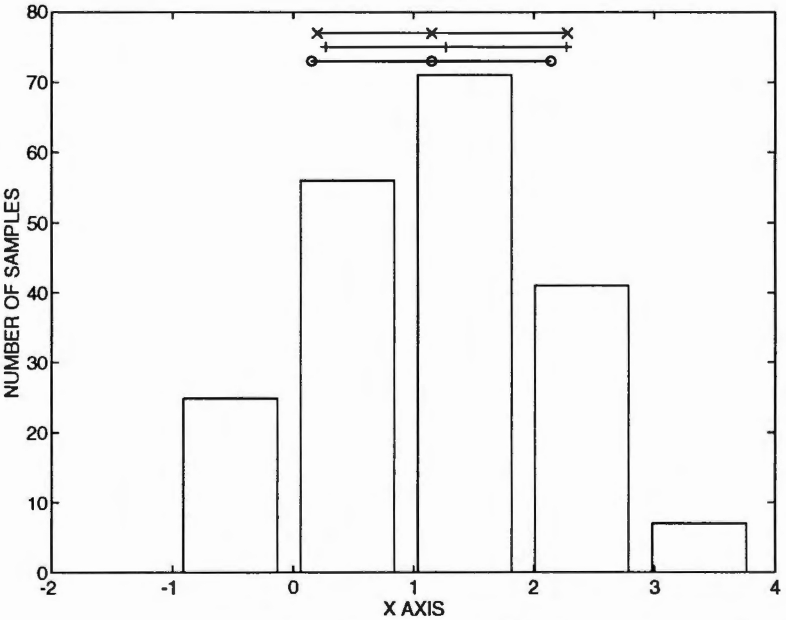


Figure 9a. True STD, Sample STD, and MAD for a $N(1.27,1)$ *Sample*

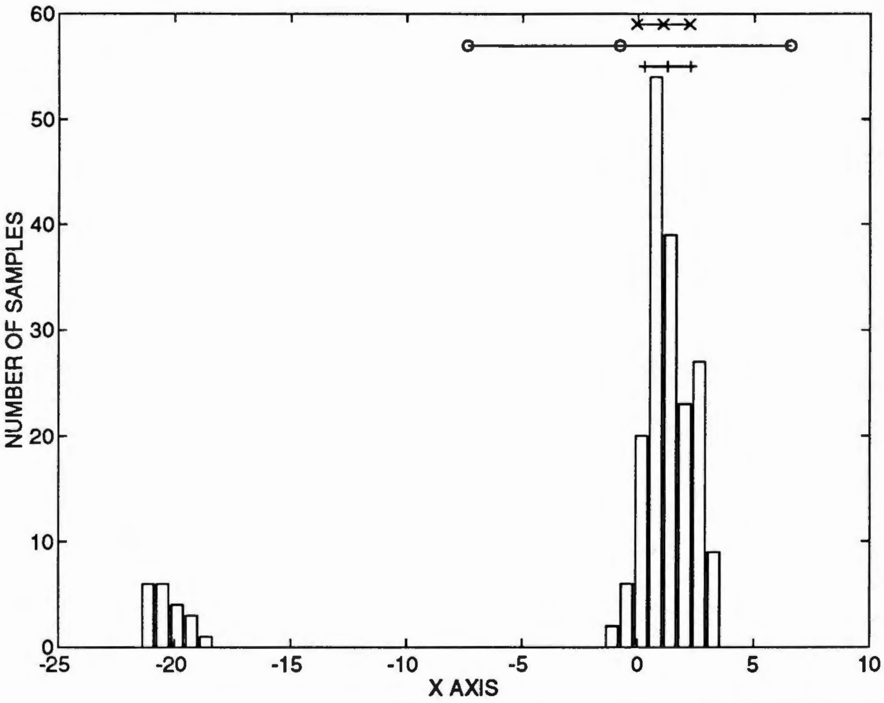


Figure 9b. True STD, Sample STD, and MAD for a Contaminated $N(1.27,1)$ ***Sample***

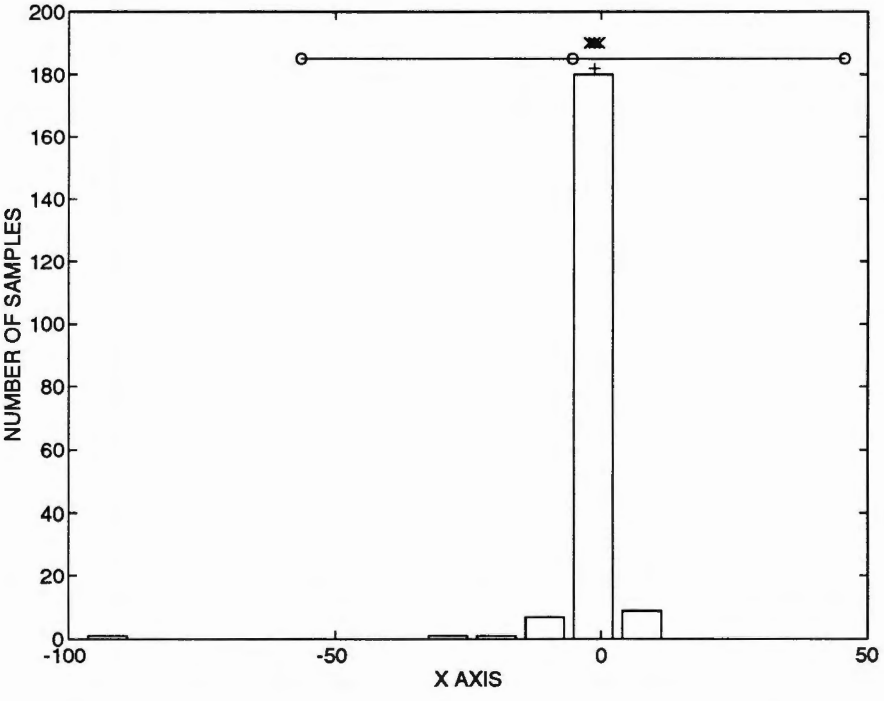


Figure 9c. True STD, Sample STD, and MAD for a Cauchy Sample

The previous examples compared two of the most vulnerable statistics, the sample central moments, against the two most resistant statistics, the median and the MAD. However, both of these latter statistics tend to ignore magnitude information in the sample values. One can see this by realizing that the median is the solution to the equation

$$\sum_{k=1}^{N} sgn(X_k - c) = 0 \ ,$$

where c is the resulting median. Since *no* magnitude information from the sample is used in this equation, the median is more resistant to outliers. In contrast, *all* the magnitude information is used in the sample mean since the IF is the identity function. Moreover, the linearity of the sample mean implies that any outlier in the sample has the same impact as any other sample, thus destroying the estimate. In between these two extremes are the M-estimators of Huber, which tend to use the magnitude information from the samples, but only in a region where the samples are not likely to be outliers. The covariance estimator of this memorandum is an M-estimator based on Huber's work.

So far, only one-dimensional examples have been discussed. To apply these median and MAD estimators to vector samples, the estimators are applied on a component-by-component basis. The covariance estimate, however, is more sophisticated than just a diagonal estimate of the covariance matrix where the diagonal elements are the square of the MAD estimates made on the component-by-component basis. The diagonal estimate can be used for an initial estimate of the iterative estimation procedure. Instead, an M-estimator based on the radial distance from the center of the sample is used to reduce the influence of the outliers on the dispersion estimate. In the next section, the robust covariance estimate is explained in detail.

4.3 ROBUST COVARIANCE ESTIMATION ALGORITHM

The covariance estimation algorithm is based on an algorithm found in Huber's text (reference 5, p. 215-221 and p. 237-242). This algorithm is not transparent since the basic idea has been altered to obtain numerical stability. First, assume that the centering vector is zero. The easiest way to envision the algorithm is as a fixed point minimization problem. The recursion for the covariance matrix has the following form:

$$\hat{\Sigma}_{(m+1)} = \frac{avg(s(d_m)xx^t)}{avg(s(d_m))} \ ,$$

where

$$d_m(x) = x^t \hat{\Sigma}^{-1}_{(m)} \, x$$

is the normalized distance at step m based on the previous estimate of the covariance matrix. Note that the *avg* operator produces the same effect as a sum operator since the averaging over the number of samples cancels out on top and bottom. This recursion shows that the estimate is a recursive weighted minimization procedure much like the reweighted sum of squares method. The weight functions $s(\cdot)$ are maximum near the centering vector and drop off with the distance from the center, thus de-emphasizing the outliers. Very specific forms for the weight functions $s(\cdot)$ are required to ensure convergence. In short, this procedure is a modified rejection procedure, where instead of rejecting outliers, one de-emphasizes them gradually.

For reasons of stability, Huber does not suggest this algorithm be used in actual computation. Instead, he writes $\hat{\Sigma}_{(m)} = (V^t_{(m)} V_{(m)})^{-1}$ and $\Sigma^{-1} = V^t V$ where $V = B^{-1}$,

$\Sigma = BB^t$, and B is the Choleski decomposition of the covariance matrix. Then, $d_m(x) = \left|V_{(m)}\, x\right|$, where the squared Euclidean distance is defined by $|x| = \sqrt{x^t x}$. The algorithm given here assumes that the measure of central tendency is fixed as the median vector. This procedure is a modification of a more general algorithm developed by Huber (reference 5, p. 238). The first step and perhaps the most important is the initial estimate of the centering vector and covariance matrix. This issue will be addressed separately and for now these two parameters are just assumed.

1. Initialize the centering vector and covariance matrix: $t := t_0$, $\Sigma := \Sigma_0$ and calculate V.

2. Construct estimate of the covariance matrix. From the data, calculate a transformed data set $y = V(x - t)$ and set

$$C := \frac{avg(s(|y|)yy^t)}{avg(s(|y|))}\text{, or equivalently, } C := \frac{\sum_{i=1}^{N}(s(|y_i|)y_i\, y_i^t)}{\sum_{i=1}^{N} s(|y_i|)}$$

from which is formed the Choleski decomposition where $C = BB^t$. Set $W := B^{-1}$ and $V := WV$.

3. Stopping Rule: STOP if the norm of W is below some prescribed level. More explicitly, if $\|W - I\| < \varepsilon$, where ε is application dependent, then STOP, else, go to step 2. There are many matrix norms from which to choose; however, the maximum absolute row sum was initially chosen to implement in code.

This algorithm has been implemented in C using Recipes in C for the numerical subroutines. The Choleski decomposition algorithm was obtained from a classical numerical analysis text (reference 10). Intuitively, this algorithm iteratively centers and whitens the data. The centering vector is a constant in this version, so in step 2, the centered data are whitened by the matrix V and then the whitened data are used to estimate a covariance matrix, taking into account the new weights for discounting the outliers. The new estimator of the covariance matrix is decomposed using the Choleski decomposition and a new whitening matrix is formed from the product of the old whitening matrix and a multiplicative correction matrix. In the limit, $W \rightarrow I$, $C \rightarrow I$, and V now has approached the "square root" of the covariance matrix. Since this is basically an accelerated fixed point algorithm, the initial starting point is very important. The initialization is discussed in the next section.

The weight functions $s(\cdot)$ are not arbitrary; there are several conditions that must be satisfied by these functions (references 3, 5, 9). The function chosen for this memorandum is

$$s(\cdot) = \begin{cases} 1, & |x| \le 1 \\ 1/|x|, & |x| > 1 \end{cases},$$

which down weights the outliers far removed from the centroid or centering constant of the cluster.

4.4 ROBUST INITIAL ESTIMATES OF THE COVARIANCE MATRIX

Since the covariance estimation technique is iterative and nonlinear, it requires a good initial estimate to assure its efficiency and convergence. Three different initialization techniques have been developed. The first initial estimator of the covariance is a modified form of the sample covariance, where the centering constant has been replaced by the median. The second is a brute force technique where the covariance matrix is initialized as the identity matrix or as a diagonal matrix, $\hat{\Sigma}_{(0)} = diag(mad^2(X_1),\cdots,mad^2(X_N))$ whose diagonal entries are the square of the MAD estimates of the dispersion. Both of these approaches produce a positive definite matrix for the initial covariance matrix; however, both estimates ignore the correlation. The last technique, which was suggested by Huber, uses variance estimates to construct the correlation estimates (reference 5, p. 202). When the variance is replaced by MAD estimates, one can obtain a reasonable estimate for the covariance matrix except that the matrix is not guaranteed to be positive definite. A way around this deficiency is to find the eigenvalues of the robust estimate, replace any negative eigenvalues with small positive eigenvalues, and then reconstruct the initial estimate of the covariance. Since the first two covariance matrix estimates are straight-forward, only the Huber covariance estimate is discussed in detail.

A robust initial estimate of the covariance matrix can be achieved through the observation that the covariance can be written as the difference of two variances (reference 5). Thus, the robust estimates of scale can be used to construct robust estimates of the covariance:

$$\mathrm{cov}(X,Y) = \frac{1}{4}[\mathrm{var}(X+Y) - \mathrm{var}(X-Y)],$$

which appears to be a trivial identity except when coupled to the fact that there are powerful robust estimates of variance. This allows us to construct a robust initial estimate of covariance. Although there are other robust estimates of covariance such as the Spearman's rank correlation and Kendall's Tau coefficient, the above technique allows us to extend robust estimates of dispersion to robust estimates of correlation.

The MAD estimator can be used to construct robust estimates of the covariances. In particular, one estimate of the correlation coefficient for two random variables is given by

$$\rho_{12} = \frac{\mathrm{var}(X_1+X_2) - \mathrm{var}(X_1-X_2)}{\mathrm{var}(X_1+X_2) + \mathrm{var}(X_1-X_2)},$$

which is now easily robustified by replacing the variance estimates by robust measure of dispersion. The covariance terms are then

$$\mathrm{cov}(X_1,X_2) = std(X_1)\rho_{12}std(X_2),$$

where std is the STD or the square root of the variance. The robustified estimate, denoted by robcovar, is

$$robcovar(X_i,X_j) = mad(X_i)\frac{mad^2(X_i+X_j) - mad^2(X_i-X_j)}{mad^2(X_i+X_j) + mad^2(X_i-X_j)}mad(X_j),$$

where $mad(X_i + X_j)$ is the MAD estimate of the random variables $\left\{X_i + X_j\right\}$, or the sum of the i-th and the j-th components of the random data vectors. The covariance matrix is then $\Sigma = \left[\sigma_i \rho_{ij} \sigma_j\right]$, whereas the robust estimate of the covariance matrix is

$$rob\,\Sigma = robcovar(X_i, X_j).$$

Unfortunately, robcovar is not necessarily a covariance matrix; that is, it need not be positive definite. Therefore, it might need to be modified if it is to be used as the start of a recursive estimation procedure for the covariance matrix. To see this, the general form of the recursive estimator must be considered. The recursion is

$$\Sigma_{(m+1)} = \frac{\sum_{k=1}^{N} u(x_k^t \Sigma_{(m)}^{-1} x_k) x_k x_k^t}{\sum_{k=1}^{N} u(x_k^t \Sigma_{(m)}^{-1} x_k)},$$

where $x_k x_k^t$ is the outer product of the k-th random sample vector and $x_k^t \Sigma_{(m)}^{-1} x_k$ is the distance from the centering vector, assumed to be zero here. For $d(x_k) = x_k^t \Sigma_{(m)}^{-1} x_k$ to be a metric, both Σ and its inverse must be positive definite (reference 12, p.393). Thus, some modification of the initial covariance must be made to be sure that the metric makes sense. In practice, one may wish to ignore this, knowing that on the next step the covariance estimate formed from the recursive equation will be positive definite.

A viable alternative to obtain a positive definite estimate of the covariance matrix is simply to modify the covariance matrix to force its positive definiteness. Since the estimate $\Sigma_{(0)}$ is surely symmetric, its eigenvalues are real. Moreover, its eigenvector matrix is orthogonal, allowing the standard similarity transformation (reference 11, p. 312) $\Phi^t \Sigma \Phi = \Lambda$, where Φ is the eigenvector matrix and Λ is the diagonal matrix of eigenvalues. If all the eigenvalues of Σ are strictly positive, then one can construct the Choleski decomposition directly from Σ. Else there are some eigenvalues that are zero or negative. Replace these latter eigenvalues with some small positive value δ and then obtain $\tilde{\Lambda}$ as the modified version of Λ. Then, the initial estimate can be obtained by setting $\hat{\Sigma}_{(0)} = \Phi \tilde{\Lambda} \Phi^t$, which is now a robustified positive definite initial estimate of the covariance matrix.

5. ROBUST CLASSIFIER DESIGN RESULTS

A comparison of the training parameter estimation is made for two distributions: the contaminated normal and the Cauchy distribution. In section 3, the vulnerability of the sample statistics employed in constructing the QC from training data was illustrated. Although the contaminated normal data exhibited only degraded performance, the Cauchy data exhibited catastrophic degradation. Robust estimators degrade more gracefully, so it is expected that the QC's built on them will degrade gracefully as well. In this section, it is shown that the robust covariance estimation technique in conjunction with the median vector, provides a more stable estimate of the classifier parameters and a thus a more stable classifier. The probability of error for all the examples is summarized in the table at the end of this section and clearly shows a significant improvement of robust statistics over the sample statistics.

The first example is the contaminated normal distribution. In this example, section 3.1 demonstrated that although the sample statistics provided reasonable behavior for the training data, the percentage of error was 12.75 compared with the theoretically expected error of 11%. The QC generated was clearly not optimal. However, using the robust covariance and median to generate the parameters from the training data, one is able to construct a better QC. The covariances for both the sample and the robust statistics are given in table 3. It is apparent that the outliers have influenced the sample statistics far more than the robust statistics. This is reflected in figures 10a and 10b, where it is clear that the robust training has resulted in a better, although still not optimal QC. Comparison with the corresponding figures of section 3.1 will convince the reader that the robust classifier looks more like the ideal discriminant. The performance improved to 10.25% error, which looks more like the theoretical ideal.

Table 3. QC Parameter Estimates, Sample vs Robust Estimates, Contaminated Normal Data

Sample mean vector for the class 2 cluster.					Sample median vector for the class 2 cluster.				
-1.161	0.007	-0.016	0.005	0.068	-1.321	0.027	-0.058	0.008	0.110
Sample covariance matrix for the class 2 cluster.					Robust covariance matrix for the class 2 cluster.				
0.963	0.101	0.125	0.035	-0.007	0.823	0.108	0.086	0.021	0.003
0.101	0.898	0.004	0.003	-0.002	0.108	0.793	0.002	0.000	0.010
0.125	0.004	1.097	0.036	0.067	0.086	0.002	0.876	0.006	0.049
0.035	0.003	0.036	0.926	-0.003	0.021	0.000	0.006	0.756	0.011
-0.007	-0.002	0.067	-0.003	0.957	0.003	0.010	0.049	0.011	0.826
Sample mean vector for the class 1 cluster.					Sample median vector for the class 1 cluster.				
-0.783	-0.038	0.047	-0.006	0.056	1.081	0.006	0.048	-0.005	0.108
Sample covariance matrix for the class 1 cluster.					Robust covariance matrix for the class 1 cluster.				
43.445	-0.049	-0.013	0.307	0.461	2.211	0.135	0.080	0.049	0.022
-0.049	0.814	0.018	-0.039	-0.016	0.135	0.731	0.033	-0.027	0.019
-0.013	0.018	1.094	0.004	0.084	0.080	0.033	0.912	0.007	0.074
0.307	-0.039	0.004	0.907	0.041	0.049	-0.027	0.007	0.754	0.046
0.461	-0.016	0.084	0.041	0.904	0.022	0.019	0.074	0.046	0.829

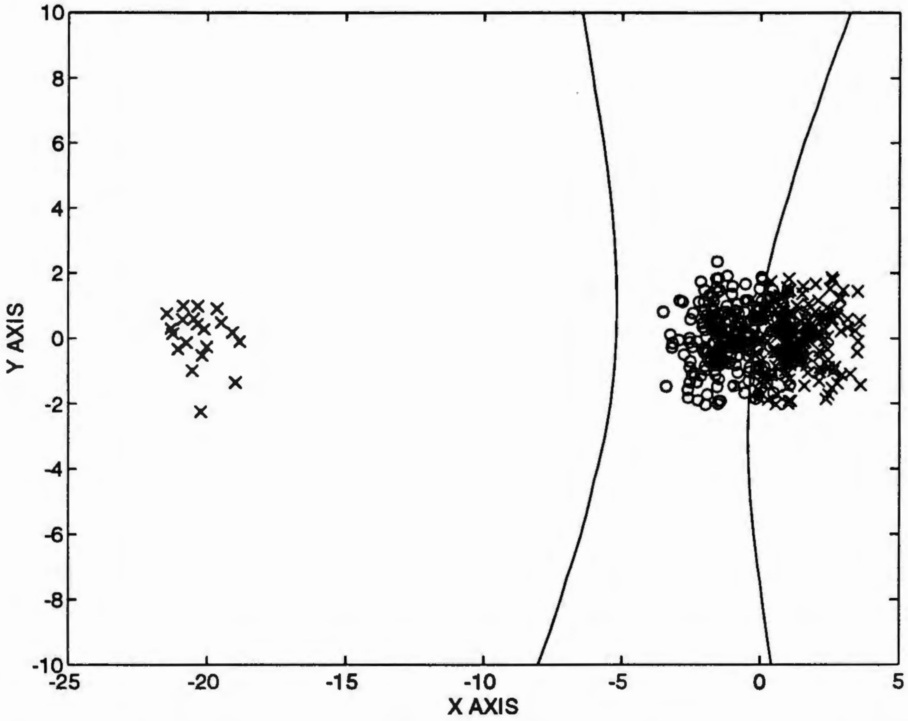


Figure 10a. QC Generated by Robust Statistics Using Normal Training Data

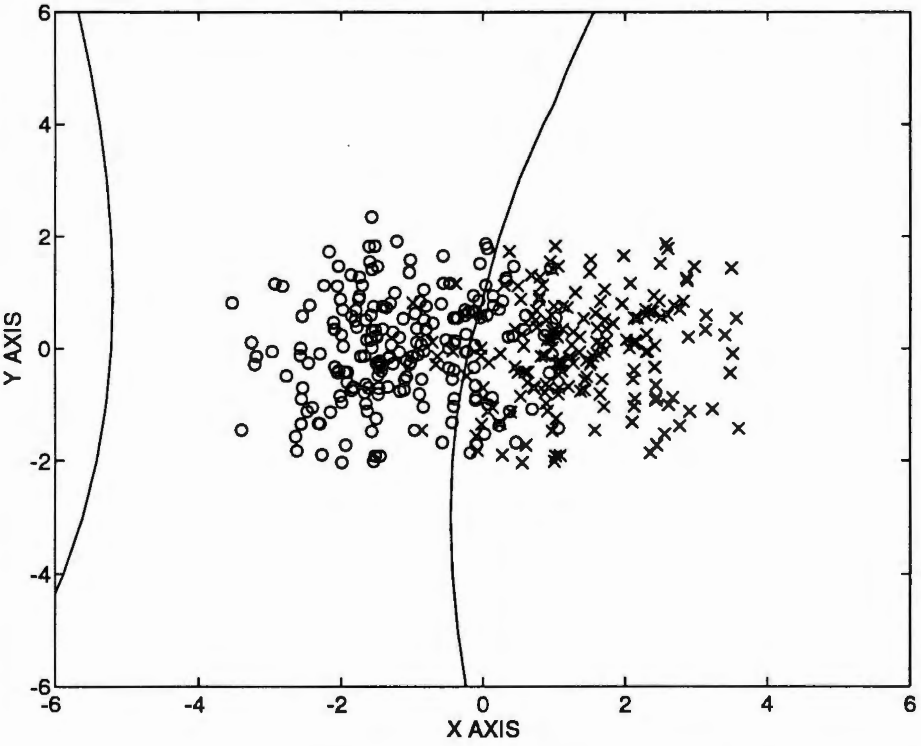


Figure 10b. QC Generated by Robust Statistics Using Normal Training Data, Expanded Scale

The second example is the Cauchy two-class problem. In section 3.2, figure 6 showed that when sample statistics were used for training the QC, the results were dismal. The error was 48.75%. Examination of the covariance estimate substantiates the deterioration of the covariance estimates. Table 4 compares the sample statistics with the robust statistics.

Table 4. QC Parameter Estimates, Sample vs Robust Statistics, Cauchy Data

Sample mean vector for the class 2 cluster.					Sample median vector for the class 2 cluster.				
-5.4	-0.0	0.0	0.1	-0.1	-1.315	-0.011	0.084	0.004	-0.069
Sample mean vector for the class 1 cluster.					Sample median vector for the class 1 cluster.				
0.9	-4.3	12.7	0.7	0.2	1.258	0.011	-0.005	0.091	-0.046
Sample covariance matrix for the class 2 cluster.					Robust covariance matrix for the class 2 cluster.				
2609.3	-8.6	-0.4	5.5	-4.7	3.036	-0.033	0.080	0.115	-0.052
-8.6	138.8	0.7	0.1	-0.3	-0.033	1.489	0.021	-0.057	-0.145
-0.4	0.7	26.4	0.3	-1.3	0.080	0.021	2.197	0.033	-0.054
5.5	0.1	0.3	26.2	-0.2	0.115	-0.057	0.033	1.745	-0.110
-4.7	-0.3	-1.3	-0.2	52.1	-0.052	-0.145	-0.054	-0.110	1.998
Sample covariance matrix for the class 1 cluster.					Robust covariance matrix for the class 1 cluster.				
12.0	-3.8	-3.5	1.0	0.7	1.268	-0.052	-0.021	0.029	0.070
-3.8	5246.2	65.7	-11.4	-13.8	-0.052	1.421	-0.118	-0.078	-0.004
-3.5	65.7	13942.4	-7.8	1.0	-0.021	-0.118	1.449	0.039	-0.042
1.0	-11.4	-7.8	53.5	-0.8	0.029	-0.078	0.039	2.022	-0.020
0.7	-13.8	1.0	-0.8	11.1	0.070	-0.004	-0.042	-0.020	1.423

The heavy-tailed Cauchy distribution is often used to represent the impact of outliers on the covariance estimate. Although few in number, the extreme outliers have given a poor estimate of the covariance matrix and this in turn has destroyed the covariance estimate. However, the robust covariance estimate has given reasonable values and the QC that it yields is shown in figure 11. In this figure, the ideal classifier is the y-axis with samples falling to the left and right of this boundary classified into separate classes. The quadratic nature of the classifier in this figure is illustrated in its elliptic shape, with samples within the ellipse falling into one class and samples outside the ellipse falling into the other class. The orientation and eccentricity of the ellipse provide a good boundary approximation to the y-axis.

Table 5. Probability of Error

Data\Estimator	Ideal	Sample	Robust
Normal	13.5%	12.75%	12.25%
Outliers in Normal	11.0%	12.75%	10.25%
Cauchy	10.8%	48.75%	20.25%

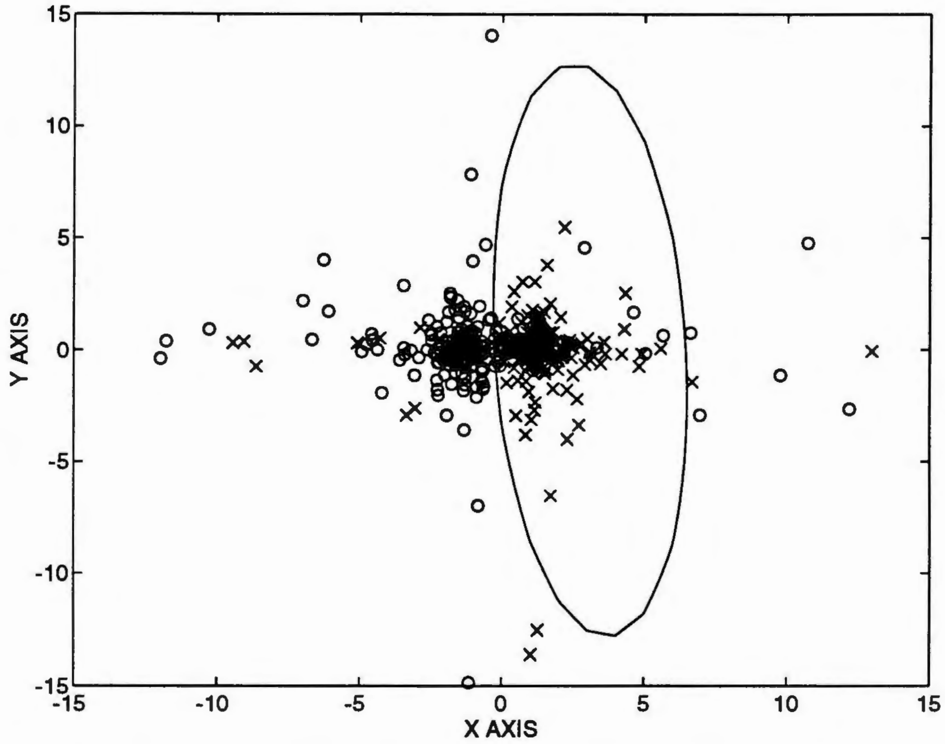


Figure 11. QC Generated by Robust Statistics Using Cauchy Training Data

6 SUMMARY, CONCLUSIONS, AND FUTURE DIRECTIONS

In this memorandum, robust statistics were applied to the training of the quadratic classifier (QC). What is meant by training is to estimate the centering constants and the covariance matrices associated with the QC. The robust statistics are shown to be resistant to changes in the underlying noise distribution expected in the sample. Since the QC is designed assuming a normal distribution of the samples, one expects to see the best performance for the normal data. This was demonstrated in section 2. However, the sample statistics used to estimate the parameters of the QC are maximum likelihood estimators, which are sensitive to any variation of the underlying distribution. It was shown that when outliers were present or when the sample distribution was heavy-tailed, the performance of the QC designed using sample statistics degraded. In the case of the outliers, the degradation was a few percentage points for outliers on the order of 20 standard deviations from the cluster centers. However, in heavy-tailed distributions with extreme outliers, the degradation was catastrophic.

The robust estimators are based on a normal model, but are designed to perform well for variations in the underlying distribution. In section 3 a short discussion of these statistics was given along with several theoretical measures of their resistance. The breakdown point is one measure of resistance and roughly the percentage of outliers these statistics could resist before the estimates became meaningless. Sample statistics have a zero percent breakdown and the robust statistics have much higher breakdown. The performance of the robust statistical designs on the training data was shown to be close to the optimal when the distribution of the data was normal and far better when the data were either contaminated or heavy-tailed.

The recommendation is clear. Although the robust statistics are more complex and have higher time and space complexity than sample statistics, the extra investment yields high dividends in resistance to changes in the underlying distribution. In essence, one is buying insurance to protect against contaminated or very dirty training data. This will not protect against bad data when the classifier is applied for testing. It does protect against being fooled by your own training data.

The author is acutely aware that a single data set does not make a statistical study. Significant simulation studies are needed to better characterize the degradation of the QC designed with the sample statistics compared with the robust statistics. Robust statistical algorithms are non-linear recursive procedures whose convergence point can depend on the initial solution. Section 4.4 discussed three different initialization techniques for the covariance matrix. Only one was used in this memorandum, the diagonal approximation to the covariance matrix based upon the MAD estimator. The effect of different initial solutions upon the solution and the rate of convergence should be included in the study. The weight function $s(\cdot)$ was also fixed for this memo; its effect must also be studied.

Another method for estimating the parameters of the classes when viewed as clusters is based on the fuzzy c-Means clustering algorithm. Robust versions of this algorithm have already been developed by the author. These clustering algorithms need to be compared to the robust covariance estimation techniques demonstrated in this memorandum, which is only the start of a very rich research area.

REFERENCES

1. Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1990.

2. Duda O. and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, 1973.

3. Hampel, F. R., P. J. Rousseeuw, E. M. Ronchetti, and W. A. Stahel, *Robust Statistics, The Approach Based on Influence Functions,* John Wiley, New York, 1986.

4. Randles, R. H.,. and D. A.Wolfe, *Introduction to The Theory of Nonparametric Statistics*, John Wiley, New York, 1979.

5. Huber, P. J., *Robust Statistics*, John Wiley, New York, 1981.

6. Hoaglin, D. C., F. Mosteller, and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis*, John Wiley, New York, 1983.

7. Rousseeuw, P. J. and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley, New York, 1987.

8. Gibbons, J. D., *Nonparametric Statistical Inference*, McGraw-Hill, New York, 1971.

9. Maronna, A. R., "Robust M-Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 1976, Vol. 4, No. 1, p. 51-67.

10. Forsythe, G. E. and C. B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

11. Noble, B., *Applied Linear Algebra*, Second Edition, Prentice-Hall, Englewood Cliffs, NJ, 1969.

DISTRIBUTION LIST

External:
NAVSEA (ASTO-B (W. Chen), ASTO-G (G. Kamilakis), ASTO-G3 (LCDR Traweek))
ONR (Codes 4520, 4525)
NRL (Code 5510, (A. Meyrowitz))

Internal:
Codes 10
102
22
221
2211
2211 (S. Maloney)
2211 (P. Kersten (30))
38
3891
81
83
02244
0251
0261 (NLON Library)
0262 (NPT Library (2))

Total 51